# Equating Laboratories: Modelling and Analysis

David Banks and Keith Eberhardt
National Institute of Standards and Technology
Statistical Engineering Division, 8980
Gaithersburg MD, 20899 USA
banks@nist.gov, keith.eberhardt@nist.gov

September 24, 2001

## Abstract

Efficient world trade requires that manufacturers in one country have confidence that their product will meet specifications that are verified by purchasers in another country. But these trading partners rely upon different national metrology laboratories to calibrate their equipment, and there is measurable divergence between their instruments. This paper describes statistical methods that enable one to use data from a network of key comparisons to estimate the measurement functions at each participating laboratory. These methods cannot determine which laboratory is most accurate, but do allow one to predict the value that a given laboratory would obtain on an artifact from the value measured at another laboratory in the comparison network.

Introduction

International trading partners want to have confidence that their measurements on shipped goods will substantially agree. But the supplier and purchaser use different national metrology laboratories to calibrate their measurement instruments, and there is small but detectable drift across the measurement chain that connects the partners. It would be helpful if the supplier could accurately forecast the measurements that the purchaser will obtain when the delivered product is tested. To support this, we develop statistical theory to equate laboratories based upon sets of artifacts measured by two or more laboratories in the calibration chain.

Metrology laboratories regularly calibrate themselves against each other by obtaining measurements upon the same artifact. Assuming that transportation does not affect the true value for the artifact, then any two laboratories are linked by a chain (or, more accurately, a network) of key comparisons which, up to noise, should agree. These data provide the basis for a statistical analysis that can track small drifts across intermediate links, allowing the supplier to predict the values that will be measured by the purchaser.

This problem of equating laboratories can be formulated either in frequentist or Bayesian terms. Both approaches are described below, and the qualitative differences are discussed. We note that our comparison focuses on the simplest practical analysis in each case—both inferential methods can be made more complex and more realistic. We suggest that such analyses be instantiated in a software program to be called MENSOR (Measurement Equivalence from National Standards and Observed Relationships), and made available on the World-Wide Web using Java applets to ensure confidentiality about measurement capability at individual laboratories.

Frequentist Analysis

Suppose there are $I$ measurement laboratories and $J$ reference artifacts.

Artifact $j$ is measured $n_{ij}$ times by measurement laboratory $i$, producing observations $Y_{ijk}$ for

$k1, \ldots, n_{ij}$. We assume the model

$$Y_{ijk} f_i(\mu_j) + \epsilon_{ijk}$$

where $\mu_j$ is the unknown true value for the $j$th reference artifact, $f_i$ is an unknown smooth function (in a sense made precise later) for the $i$th laboratory, and the $\epsilon_{ijk}$ are independent $N(0, \sigma_{ij}^2)$ random variables (so measurement error variability depends upon both the laboratory and the artifact; this happens, for example, with chemical assays, where precision depends upon the concentration of the analyte).

One would like to estimate the measurement function $f_i$ for each laboratory, as this would enable one to estimate the true value $\mu_j$ for each artifact. But the problem is ill-posed; e.g., the $f_i$ function includes the bias, which is impossible to recover from data. Fortunately, one doesn't need to know true values in order to solve the simpler problem of predicting another laboratory's results.

The first issue is to model the $f_i$. Metrology scientists rarely have more than ten reference artifacts measured in common between any pair of laboratories. Given these small sample sizes, one cannot do more than fit a low-order polynomial, such as the simple linear regression model

$$\mathbb{E}[Y_i^*] f_i(\mu)\alpha_i + \beta_i \mu^* \tag{1}$$

where $Y_i^*$ is a measurement made at the $i$th laboratory on an artifact with true value $\mu^*$. In rare cases, one might have enough data to warrant second-order polynomial regression, fitting the model

$$\mathbb{E}[Y_i^*] f_i(\mu)\alpha_i + \beta_i \mu^* + \gamma_i {\mu^*}^2.$$

Higher-order polynomials require large numbers of reference artifacts. Forecasts for new artifacts whose values lie outside the range of the reference artifacts will have large uncertainties.

As an alternative to linear or polynomial fitting, one could use nonparametric regression, but this requires enormous numbers of reference artifacts to be measured in common. Kolen and Brennan [4] describe such work in the context of educational testing—here the reference artifacts are students, so large samples are available from test piloting exercises.

The relation in (1) is called Mandel's bundle-of-lines model [6], which is used to test for interaction in two-way layouts without replicated observations (it generalizes Tukey's one-degree-of-freedom test for nonadditivity). Our application is somewhat different, in that:

1. Many cells in the two-way layout have no observations (i.e., there are incomplete blocks).

2. When a cell does contain an observation, that measurement is usually replicated.

We use the bundle-of-lines model to estimate relations between the approximate measurement functions at the different laboratories, instead of as a test for interaction.

One consequence of use of the bundle-of-lines model in our application is that some parameters are not estimable. In particular, one can neither estimate the true value $\mu_j$ of an artifact nor the linear function for the $i$th laboratory. However, one can usually estimate contrasts between, say, the manufacturer's laboratory and the purchaser's laboratory, which is sufficient for forecasting purposes. The only case in which contrast estimates cannot be made is when there is no chain of laboratories, linked by key comparisons on common reference artifacts, between the two laboratories. Usually, there is not merely a chain, but rather a network of measurements, and the bundle-of-lines model automatically pools information from all possible paths linking one laboratory to another.

Frequentist Analysis: Example

Because of the product term $\beta_i \mu_j$ in (1), Mandel's bundle-of-lines model is not a traditional linear model—the product prevents representation of the measurement as a simple sum of unknowns. Therefore the model

$$Y_{ijk}\alpha_i + \beta_i \mu_j + \epsilon_{ijk}$$

is traditionally estimated in two steps.

To illustrate the details, consider a two-way layout for four laboratories and eight reference artifacts, with two replications for each measurement. The example in Table 1 was constructed by taking the index

of the artifact as the value of the artifact (thus the $j$th artifact has value $j$), and the measurement functions for the four laboratories as $f_1(\mu)1+\mu$, $f_2(\mu)2\mu$, $f_3(\mu)3+\mu$, and $f_4(\mu)-1+3\mu$. Measurement errors are independent $N(0,.01)$ random variables. Although no reference artifacts are measured in common by Laboratories 1 and 4, one might want to predict the measurement Laboratory 4 would obtain on an artifact from the value found at Laboratory 1.

| Artifacts | Laboratories | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| 1 | (2)<br>2.2<br>1.9 | (2)<br>1.9<br>2.1 | | |
| 2 | (3)<br>3.0<br>3.0 | (4)<br>4.1<br>4.1 | | |
| 3 | (4)<br>3.9<br>4.0 | | (6)<br>5.9<br>6.0 | |
| 4 | (5)<br>5.0<br>5.1 | | (7)<br>7.0<br>6.9 | |
| 5 | | (10)<br>10.0<br>10.0 | | (14)<br>13.9<br>13.9 |
| 6 | | (12)<br>11.9<br>12.0 | | (17)<br>17.1<br>17.0 |
| 7 | | | (10)<br>10.1<br>10.1 | (20)<br>19.9<br>20.1 |
| 8 | | | (11)<br>11.0<br>11.1 | (23)<br>23.0<br>23.1 |

Table 1: Example of key comparison data.

Following Milliken and Johnson [6], the two steps in estimating the parameters for a bundle-of-lines model are to:

1. fit the additive model $\mathbb{E}[Y_{ijk}]m+\tau_i+\gamma_j$.
2. calculate the residuals $r_{ijk}$ from the additive model, and fit the model $r_{ijk}\delta_i\hat{\gamma}_j+\epsilon_{ijk}$.

Some additional calculation is needed to obtain the most useful final form, as written in (1).

Applying this algorithm to the data in Table 1, the first step produces (among other output) the estimated $\mu_j$ values and associated uncertainties. These estimates are not unique—the values shown are obtained under the constraint that the largest $\mu_j$ is set to zero. SAS finds the estimated artifact effects are $\hat{\mu}_1 - 11.800$, $\hat{\mu}_2 - 10.275$, $\hat{\mu}_3 - 7.619$, $\hat{\mu}_4 - 6.569$, $\hat{\mu}_5 - 6.356$, $\hat{\mu}_6 - 3.806$, $\hat{\mu}_7 - 2.000$, and (by constraint) $\hat{\mu}_8 0.000$. These estimates are recentered about their mean, then used as explanatory variables in the second step which fits the model (1) to produce estimates of the laboratory intercepts and slopes, with associated uncertainties. These estimates are also non-unique, since they depend upon the artifact effects.

In examining the results, it is clear that the fitted model provides poor description, and is especially deficient in forecasting results for the unobserved combinations of laboratories and artifacts. This is caused by the incomplete block design, which confounds the artifact effect with the interaction (product term), preventing proper extraction of the $\hat{\mu}_j$ terms needed for regression on the residuals.

The problem lies in the fitting algorithm, not in the model. To avoid the difficulty, we use direct least squares optimization to find the coefficients for the $f_i$ in the Mandel bundle-of-lines model. Additionally, one wants confidence bounds on these estimates, and for that we use the jackknife technique (cf. Efron [1]). This is an approximate method, and tighter bounds could be obtained with somewhat more work. The jackknife estimate of standard error is

$$\hat{\mathrm{Var}}[\hat{Y}_{ij}]\frac{n-1}{n}\sum_{i',j',k}(\hat{Y}_{(i',j',k)}-\hat{Y}_{ij})^2$$

where $\hat{Y}_{ij}$ is the estimate obtained from the least squares optimization using the entire dataset, and $\hat{Y}_{(i',j',k)}$ is the estimate obtained by applying least squares optimization to all of the dataset except for observation $k$ from laboratory $i'$ on artifact $j'$. Es-

sentially, we use the average instability in the prediction, under deletion of each observation in turn, as an estimate of the variance in the forecast. Since, by design, certain observations are highly influential, this approach overstates the real uncertainty.

| | Laboratories | | | |
|---|---|---|---|---|
| Artifacts | 1 | 2 | 3 | 4 |
| | (2) | (2) | (4) | (2) |
| 1 | 2.01 | 2.02 | 3.91 | 0.95 |
| L | 1.71 | 1.76 | 3.42 | -1.72 |
| U | 2.31 | 2.29 | 4.40 | 3.62 |
| | (3) | (4) | (5) | (5) |
| 2 | 3.06 | 4.07 | 4.99 | 4.27 |
| L | 2.93 | 3.99 | 4.62 | 2.23 |
| U | 3.19 | 4.15 | 5.36 | 6.32 |
| | (4) | (6) | (6) | (8) |
| 3 | 3.98 | 5.86 | 5.93 | 7.18 |
| L | 3.86 | 5.37 | 5.79 | 5.40 |
| U | 4.10 | 6.36 | 6.06 | 8.95 |
| | (5) | (8) | (7) | (11) |
| 4 | 5.01 | 7.89 | 6.99 | 10.46 |
| L | 4.86 | 7.26 | 6.84 | 8.94 |
| U | 5.16 | 8.52 | 7.13 | 11.98 |
| | (6) | (10) | (8) | (14) |
| 5 | 6.10 | 10.01 | 8.10 | 13.90 |
| L | 5.47 | 9.89 | 7.79 | 13.81 |
| U | 6.72 | 10.12 | 8.41 | 13.98 |
| | (7) | (12) | (9) | (17) |
| 6 | 7.10 | 11.95 | 9.12 | 17.05 |
| L | 6.28 | 11.78 | 8.88 | 16.88 |
| U | 7.91 | 12.12 | 9.36 | 17.22 |
| | (8) | (14) | (10) | (20) |
| 7 | 8.03 | 13.78 | 10.08 | 20.01 |
| L | 7.03 | 13.12 | 10.00 | 19.75 |
| U | 9.03 | 14.43 | 10.16 | 20.27 |
| | (9) | (16) | (11) | (23) |
| 8 | 8.99 | 15.65 | 11.06 | 23.05 |
| L | 7.76 | 14.74 | 10.97 | 22.91 |
| U | 10.22 | 16.56 | 11.15 | 23.19 |

Table 2: Estimates from the bundle-of-lines model.

The results of these calculations for the data in Table 1 are shown in Table 2. The parenthetical entry in each cell is the true value, the next entry is the predicted value, and the following two entries are lower and upper 95% confidence bounds on the predicted value, respectively. The predicted values are very close to the true values, but the confidence bounds are often unfortunately wide, especially in laboratory-artifact combinations for which no observations are available.

To improve the uncertainty statements, one could use a linearization argument (i.e., Taylor's theorem in the traditional propagation of error technique), or make direct assumptions on the structure of the error covariance matrix. In this case, it is reasonable and realistic to exploit the fact that all measurement errors have the same (unknown) standard deviation, which is not used in the jackknife calculation, but

which could be easily implemented in a parametric bootstrap analysis (cf. Efron [2]).

Bayesian Analysis

A Bayesian analysis requires one to place a prior distribution over the unknown parameters. This prior should incorporate expert knowledge, but not force the outcome. In that spirit, we tend to use vague priors (e.g., maximum entropy priors, as advocated by Weise and Wöger [8] for metrology applications), but recognize that high precision metrology can take honest advantage of unusually concentrated error distributions and prior physical knowledge.

For the Bayesian formulation, it is convenient to rewrite (1) in matrix form:

$$\boldsymbol{Y}\,\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Here $\boldsymbol{Y}$ is the $n \times 1$ vector of observations $Y_{ijk}$, for $nn_{\cdot\cdot}$, the number of measurements at all laboratories on all reference artifacts; $\boldsymbol{X}$ is the $n \times 2I$ design matrix whose row corresponding to observation $Y_{ijk} contains zeroes except for the 2i-1 st and 2i th columns, which are 1 and \mu_j$, respectively; and $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of measurement errors.

A Bayesian analysis requires as input prior beliefs about the distributions

of the unknown quantities—usually this is obtained from experts. In this application, we use the following notation and models for those prior beliefs:

1. The prior on $\boldsymbol{X}$ is denoted by $\pi(\boldsymbol{X})$. This is a degenerate distribution, since only the $\mu_j$ values of the artifacts (which appear in the even-numbered columns of the $\boldsymbol{X}$ matrix) are unknown. Thus we assume that $\pi(\boldsymbol{X})$ is a matrix-valued normal distribution that is entirely specified by the random vector $\boldsymbol{\mu} \sim N(\boldsymbol{m}, \boldsymbol{D})$, a non-degenerate normal distribution. It is reasonable to assume that the covariance matrix $\boldsymbol{D}$ is diagonal, since the true values of the artifacts should be independent of one another. The expert's opinion determines $\boldsymbol{m}$ and the the diagonal entries of $\boldsymbol{D}$.

2. The prior on $\boldsymbol{\beta}$ is denoted by $\pi(\boldsymbol{\beta})$. It is usually reasonable to take $\pi(\boldsymbol{\beta})N(\boldsymbol{b}, \boldsymbol{B})$. In most metrology applications, an expert would assume $\boldsymbol{b}(0,1,0,1,\ldots,0,1)^{\mathrm{T}}$, since the intercept and slope in the measurement function for each competent laboratory should be very close to 0 and 1, respectively. Similarly, the covariance matrix $\boldsymbol{B}$ would probably have the block-diagonal form $\tau\,\mathrm{diag}(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_I)$, so that the judgments on slope and intercepts are independent for different laboratories, but scaled by an overall uncertainty captured by $\tau$. (In a more athletic analysis, one would place a hyperprior over $\tau$, but that level of detail is not crucial.)

3. The distribution of the error is denoted by $\psi(\boldsymbol{\epsilon})$. Usually one would model this as a $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ distribution,

   where the mean indicates unbiased errors and $\boldsymbol{\Sigma}$ is diagonal or block-diagonal, probably scaled by an overall uncertainty $\sigma$ as above. The covariance matrix $\boldsymbol{\Sigma}$ can and should model magnitude-dependent precision, or correlation among repeated measurements on the same artifact at the same laboratory.

It is realistic to assume that each of the unknowns is independent of the others, provided that the prior on $\boldsymbol{\epsilon}$ properly models laboratory and artifact effects (i.e., site-dependent or magnitude-dependent precision, which can be well-captured by $\boldsymbol{\Sigma}$).

The usual Bayesian linear model (cf. Lindley and Smith [5]) takes the $\boldsymbol{X}$ matrix to be measured without error, and assumes that only the $\boldsymbol{\beta}$ vector requires a prior. But in metrology, both $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are unknown, and require explicit priors. This leads to a high-dimensional integral whose solution is usually impractical. But two alternatives are available; the first relies upon Gibbs sampling, and the second uses a conjugate normal approximation.

Bayesian Analysis: Example

Suppose measurement $Y_i^*$ is obtained on an object at laboratory $i$, and one wants to predict the value that laboratory $i'$ will obtain when the same object is measured there. To support that inference, one can use the data on the network of reference artifact comparisons shown in Table 1.

**Exact Bayesian Solution**

The first step is to find the posterior distribution

of $\boldsymbol{\beta}$. The second step is to find the posterior distribution of $\mu^*$, the true value of the object, given the measurement $Y_i^*$. The third step is to find the posterior distribution of the measurement that will be obtained at laboratory $i'$, given $\mu^*$.

Exact Bayesian analysis first finds the distribution of $\boldsymbol{\beta}$ given the observations $\boldsymbol{Y}$ (in our notation, this distribution is $\pi(\boldsymbol{\beta} \mid \boldsymbol{Y})$, where the bar indicates that the distribution of the first argument is conditioned on the value of the second). ¿From the calculus of conditional probabilities,

$$\pi(\boldsymbol{\beta} \mid \boldsymbol{Y}) \int \pi(\boldsymbol{\beta}, \boldsymbol{X} \mid \boldsymbol{Y})\, d\boldsymbol{X}$$

where Bayes' theorem implies

$$\begin{aligned} \pi(\boldsymbol{\beta}, \boldsymbol{X} | \boldsymbol{Y}) &\propto g(\boldsymbol{Y} | \boldsymbol{\beta}, \boldsymbol{X}) \pi(\boldsymbol{\beta}, \boldsymbol{X}) \\ &\propto g(\boldsymbol{Y} | \boldsymbol{\beta}, \boldsymbol{X}) \pi(\boldsymbol{\beta}) \pi(\boldsymbol{X}). \end{aligned} \quad (2)$$

Here $g(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{X})$ is the distribution of $\boldsymbol{Y}$ as a function of specific values of $\boldsymbol{X}$ and $\boldsymbol{\beta}$—under our assumptions, it is just $N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$. The proportional signs ($\propto$) are made into equalities by normalization, which requires integration of the right-hand side in (2) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{X}$. The factorization of $\pi(\boldsymbol{\beta}, \boldsymbol{X})$ as $\pi(\boldsymbol{\beta})\pi(\boldsymbol{X})$ uses the assumption that $\boldsymbol{\beta}$ and $\boldsymbol{X}$ are independent random variables. If both are, say, normal random variables, then all terms on the right-hand side are known, and one can solve for $\pi(\boldsymbol{\beta} \mid \boldsymbol{Y})$.

The second step finds the posterior distribution of $\mu^*$, given the observed value $Y_i^* y_i^*$ and the posterior distribution $\pi(\boldsymbol{\beta}|\boldsymbol{Y})$. Using Bayes' theorem,

$$\pi(\mu^* \mid y_i^*) \frac{\pi(y_i^* \mid \mu^*)\pi(\mu^*)}{\int \pi(y_i^* \mid \mu^*)\pi(\mu^*)\, d\mu^*} \quad (3)$$

where $\pi(\mu^*)$ is the prior distribution for the analyst's belief about the true value of the new object, and $\pi(y_i^* \mid \mu^*)$ is the distribution of possible observations $Y_i^*$ for a fixed value of $\mu^*$. One might reasonably assume a normal distribution for $\pi(\mu^*)$, but the distribution $\pi(y_i^* \mid \mu^*)$ is more problematic. Recall that at laboratory $i$,

$$Y_i^* \beta_{2i-1} + \beta_{2i} \mu^* + \epsilon_i^*.$$

If the joint distribution of $\beta_{2i-1}$ and $\beta_{2i}$ were normal, then calculation would be easy. But the posterior $\pi(\boldsymbol{\beta} \mid \boldsymbol{Y})$ found in (2) is not normal, and numerical solution is needed.

For the final step, one calculates $\pi(y_{i'}^* \mid y_i^*)$. Let $Y_i^*$ have marginal distribution $f(y_i^*)$, and denote the joint distribution of $Y_i^*$ and $Y_{i'}^*$ by $g(y_i^*, y_{i'}^*)$. Using the definition of conditional probability and the law of total probability gives:

$$\begin{aligned} \pi(y_{i'}^* \mid y_i^*) \qquad & \frac{g(y_i^*, y_{i'}^*)}{f(y_i^*)} \\ & \frac{1}{f(y_i^*)} \int g(y_i^*, y_{i'}^* \mid \mu^*) \pi(\mu^*)\, d\mu^*. \end{aligned} \quad (4)$$

Under our assumptions, $Y_i^*$ and $Y_{i'}^*$ are conditionally independent given $\mu^*$; thus

$$g(y_i^*, y_{i'}^* \mid \mu^*) \pi(y_i^* \mid \mu^*) \pi(y_{i'}^* \mid \mu^*).$$

Substituting this, and a regrouping of (3), into (4) gives

$$\begin{aligned} \pi(y_{i'}^* \mid y_i^*) \qquad & \frac{1}{f(y_i^*)} \int \pi(y_i^* \mid \mu^*) \pi(y_{i'}^* \mid \mu^*) \pi(\mu^*)\, d\mu^* \\ & \frac{1}{f(y_i^*)} \int \frac{\pi(\mu^* \mid y_i^*) \int \pi(y_i^* | \eta) \pi(\eta)\, d\eta}{\pi(\mu^*)} \\ & \qquad \times \pi(\mu^*)\, \pi(y_{i'}^* \mid \mu^*) d\mu^* \\ & \int \frac{\int \pi(y_i^* | \eta) \pi(\eta)\, d\eta}{f(y_i^*)} \\ & \qquad \times \pi(y_{i'}^* \mid \mu^*) \pi(\mu^* \mid y_i^*)\, d\mu^* \\ & \int \pi(y_{i'}^* \mid \mu^*) \pi(\mu^* \mid y_i^*)\, d\mu^*, \end{aligned}$$

where the last step follows from a second application of the law of total probability. The second term in the integrand is known from (3); the first term is obtained from

$$Y_{i'}^* \beta_{2i'-1} + \beta_{2i'} \mu^* + \epsilon_{i'}^*$$

using the joint distribution of $\beta_{2i'-1}$ and $\beta_{2i'}$ available from $\pi(\boldsymbol{\beta} \mid \boldsymbol{Y})$. This enables a numerical solution of the integral.

Usually one has replicated measurements on $\mu^*$ at laboratory $i$, and it is easy to extend our analysis to that case. The key practical difficulty with this approach is that numerical solutions are computer-intensive.

**Gibbs Sampling**

The numerical difficulty of exact Bayesian calculation sparked the invention of Gibbs Sampling (cf. Gelfand and Smith [3]), an approximation technique that is widely used and which takes variant forms. The large literature on this offers strategies for tailoring the approach to specific problems. In the context of this application, we focus on the main algorithm, but note that improvement is possible.

The Gibbs Sampler is an iterative technique that depends upon the ability to draw a random value from a conditional distribution. To illustrate, consider the posterior calculation in the first step of the exact Bayesian analysis. For the initialization step, set $\boldsymbol{\beta}\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\mu}\boldsymbol{\mu}^{(0)}$. It is useful to block the $2I$-component vector $\boldsymbol{\beta}$ into $I$ subvectors, $\boldsymbol{\beta}_1(\beta_1, \beta_2)', \ldots, \boldsymbol{\beta}_I(\beta_{2I-1}, \beta_{2I})'$. Then at the $t+1$th iterate, the density of

$$
\begin{array}{lll}
\boldsymbol{\beta}_1^{(t+1)} & \text{is} & h(\beta_1 \,|\, \boldsymbol{\beta}_2^{(t)}, \ldots, \boldsymbol{\beta}_I^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{Y}) \\
\boldsymbol{\beta}_2^{(t+1)} & \text{is} & h(\beta_2 \,|\, \boldsymbol{\beta}_1^{(t)}, \boldsymbol{\beta}_3^{(t)}, \ldots, \boldsymbol{\beta}_I^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{Y}) \\
\vdots & \vdots & \vdots \\
\boldsymbol{\beta}_I^{(t+1)} & \text{is} & h(\beta_I \,|\, \boldsymbol{\beta}_1^{(t)}, \ldots, \boldsymbol{\beta}_{I-1}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{Y}) \\
\mu_1^{(t+1)} & \text{is} & h(\mu_1 \,|\, \boldsymbol{\beta}^{(t)}, \mu_2, \ldots, \mu_J, \boldsymbol{Y}) \\
\mu_2^{(t+1)} & \text{is} & h(\mu_2 \,|\, \boldsymbol{\beta}^{(t)}, \mu_1, \mu_3, \ldots, \mu_J, \boldsymbol{Y}) \\
\vdots & \vdots & \vdots \\
\mu_J^{(t+1)} & \text{is} & h(\mu_J \,|\, \boldsymbol{\beta}^{(t)}, \mu_1, \ldots, \mu_{J-1}, \boldsymbol{Y})
\end{array}
$$

where $h(\cdot \,|\, \cdot)$ represents a generic conditional density function. This iterative process ensures $\boldsymbol{\beta}^{(t)}$ converges in distribution to a draw from the posterior distribution $\pi(\boldsymbol{\beta} \,|\, \boldsymbol{Y})$. Repeating this iteration generates a sample from the posterior, so one can use density estimation techniques (cf. Scott [7]) to approximate the posterior.

In our application, the modeling assumptions enable simplification of the Gibbs sampler. For example, conjugacy results of Lindley and Smith [5] imply that many of the $h(\cdot \,|\, \cdot)$ densities above are known distributions that depend on only a few conditional values.

**Conjugacy Approximation**

The third approach is to approximate the true distributions by a conjugate family of distributions for which mathematical calculation is especially simple. Usually, such approaches fail, because the necessary approximations are inaccurate. In this case, at issue is the approximation of the product $\boldsymbol{X}\boldsymbol{\beta}$, where $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are independent multivariate normal random variables. Generally, this product is not normal; however, as the variance of $\boldsymbol{X}$ and/or $\boldsymbol{\beta}$ becomes small, then the product approaches normality. Also, as the mean of $\boldsymbol{X}$ and/or $\boldsymbol{\beta}$ approaches $\mathbf{0}$, the distribution of the product becomes symmetric and unimodal, which is, broadly speaking, approximately normal.

In metrology applications, it often happens that measurement variance is small, and that the distribution of the bias has mean zero. For these reasons, it is worthwhile to explore approximations that allow analytical calculation of posterior distributions for simple conjugate families.

Discussion

This work is in its early stages, and so our conclusions are tentative. The chief point is that it is important to use the entire network of common measurements, rather than just a chain. This is not an intractable problem—several statistical analyses present themselves.

Regarding the comparison of frequentist and Bayesian formulations, there is a degree of controversy here. Some statisticians are highly partisan members of one or the other statistical schools—our own view is that both are legitimate, and we should determine which approach best serves the needs of the metrologica l community.

In this application, some of the key differences are:

1. Bayesian methods use prior beliefs about the unknown quantities; this opens the door to criticism that practitioners may be overconfident in their measurement capability. But there are well-established statistical methods to prevent or discover such subjectivity.

2. Frequentist methods are inherently unable to make statements about laboratory capability, but Bayesian analyses can. Although it is not the intent of these analyses to compare measurement accuracy, some users may try to do so.

3. The Bayesian analysis gives direct probability statements about the inference, whereas the frequentist analysis presents confidence regions. The latter are more difficult to properly interpret in practice.

Administrators and practitioners must provide input on these points as both methodologies are refined.

Another practical issue is that some of the labs in the calibration chain will want to preserve privacy regarding their metrological performance. A way to achieve this is for data on key comparisons to be kept locally, and interrogated blindly by Java applets that are agents of the MENSOR program. This is not perfectly secure—an unscrupulous mathematician could pose a series of queries that would eventually determine the hidden measurements at each site. But this would be time-consuming, difficult, and easily thwarted by small randomizations, query limits, or denial of complete information about the algorithm.

Also, in commercial applications, many contracts prescribe specific lot-acceptance plans, and MENSOR should be able to handle the common ones, e.g., (i) Upon delivery, 10 resistors will be chosen at random and tested, and their average must lie between 1.21 and 1.24 ohms; or (ii) Upon delivery 10 resistors will be chosen at random and tested, and at least 9 must be between 1.21 and 1.24 ohms. In this respect one could make (at least) four kinds of uncertainty statements: Bayesian and frequentist predictive intervals on either the actual measurements or the differences between the measurements. A particularly useful statement is, for example, "MENSOR estimates that the probability that your shipment will be accepted when measured by the purchasers according to the lot-acceptance plan you have specified is .89." This is a Bayesian predictive statement about the raw measurements.

The proposed MENSOR approach entails both political and technical components. In particular, it requires an agreement to construct a (possibly decentralized) database of key comparisons, a project that is already underway. But it offers an elegant way to support international trade while avoiding the contentious problem of declaring equivalencies among national metrology institutes.

## References


[1] B. Efron. *The Jackknife, the Bootstrap, and Other Resampling Plans.* SIAM, Philadelphia, PA, 1982.

[2] B. Efron. Bootstrap confidence intervals for a class of parametric problems. *Biomtrika*, 72:45-58, 1985.

[3] A. E. Gelfand and A. F. M. Smith. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398-409, 1990.

[4] M. J. Kolen and R. L. Brennan. *Test Equating: Methods and Practices.* Springer-Verlag, New York, NY, 1995.

[5] D. V. Lindley and A. F. M. Smith. Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:1-41, 1972.

[6] G. A. Milliken and D. E. Johnson. *Analysis of Messy Data: Volume 2, Nonreplicated Experiments.* Van Nostrand Reinhold, New York, 1989.

[7] D. W. Scott. *Multivariate Density Estimation.* Wiley, New York, N.Y., 1992.

[8] Weise, K. and Wöger, W. (1993). A Bayesian theory of measurement uncertainty. *Measurement Science and Technology*, **4**, 1-11.